DOCUMENT RESUME

ED 104 936 95 TM 004 394

AUTHOR Peng, Samuel S.
TITLE The Essence of Balancing: Adjustment of Group Effects.
INSTITUTION North Carolina Univ., Chapel Hill. L.L. Thurstone Psychometric Lab.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Apr 75
GRANT NE-G-00-3-0111
NOTE 25p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D.C., March 30-April 3, 1975)

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS Analysis of Variance; *Comparative Analysis; Cultural Differences; *Groups; *Individual Differences; Matrices; Measurement Techniques; Predictor Variables; Racial Differences; *Sampling; Socioeconomic Status; *Statistical Analysis; Statistical Bias; Test Results
IDENTIFIERS *Balancing

ABSTRACT

This paper was intended to promote a deeper understanding of a statistical method called balancing developed by National Assessment of Educational Progress. Problems in estimating main effects when populations are disproportionate, balancing solutions to these problems, methods equivalent to balancing, interpretation of balanced results, and some applications are considered and accompanied with examples. It is concluded that properly balanced results or the adjusted marginal means should be considered in studies in which group status or group comparisons are of a major concern. The process of balancing can be used to identify variables relating to outcome measures, and to test for spuriousness of the group effects. (Author)

Division D/SIG/Educational Statistics
Session 7.04

ED104936

The Essence of Balancing: Adjustment of Group Effects

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRO
DUCED EXACTLY AS RECEIVED FROM
THE PERSON OR ORGANIZATION ORIGIN
ATING IT. POINTS OF VIEW OR OPINIONS
STATED DO NOT NECESSARILY REPRE
SENT OFFICIAL NATIONAL INSTITUTE OF
EDUCATION POSITION OR POLICY

Samuel S. Peng

Center for Educational Research and Evaluation
Research Triangle Institute

TM 004 394

Paper presented at the AERA Annual Meeting in Washington, D. C., April, 1975.

Abstract

This paper was intended to promote a deeper understanding of a statistical method called balancing developed by NAEP. Problems in estimating main effects when populations are disproportionate, balancing solutions to these problems, methods equivalent to balancing, interpretation of balanced results, and some applications are considered and accompanied with examples. It is concluded that properly balanced results or the adjusted marginal means should be considered in studies in which group status or group comparisons are of a major concern. The process of balancing can be used to identify variables relating to outcome measures, and to test for spuriousness of the group effects.

THE ESSENCE OF BALANCING: ADJUSTMENT OF GROUP EFFECTS[1/]

In recent years, balancing, a method developed by specialists in the National Assessment of Educational Progress (NAEP), has increasingly become an interesting topic among educational researchers and evaluation specialists. Its primary intention was to estimate group differences in the absence of masquerading by other factors. It is generally conceded that "the balanced results do a much better job than the unadjusted results of reflecting such differences" (NAEP, 1971, p. 1).

While the method seems appealing to many researchers, the only major reference currently available for the use of the method is contained in a ten-page appendix of illustrative examples (NAEP, 1971, pp. B-1 - B-10). Many questions, including the justifications for balancing, procedures, equivalent methods, interpretations of balanced results, and its applications, are not well explicated elsewhere. A better and systematic understanding of these problems seems necessary to bring forth the method's potentials and to ensure its proper applications. Therefore, the primary purpose of this paper is to synthesize available information in order to achieve a deeper understanding of balancing.

A. The Problems and Justifications for Balancing

In many educational evaluation studies, it is often necessary to make comparisons of group effects. For example, in a statewide assessment of educational progress, educators wish to know how children from different types of communities with different ethnic backgrounds differ in their knowledge in subject areas such as literature, science, and mathematics.

A plausible approach is to compute simple group means from the observed data, and then compare the means. More specifically, if one is interested in knowing, for instance, how Black children differ from White children in the proportion passing a mathematics test, one could simply compute the difference between these two proportions. This approach is simple and straightforward but may be misleading under certain circumstances. We might reasonably expect children from rural and inner city areas to perform at lower-than-average levels for a variety of reasons. It is also clear that parental education will have a substantial bearing on performance on achievement tests. If we classify children jointly on the basis of race, parental education and type of community, Blacks will largely fall into the category of inner city or rural community with low parental education. If we find a deficit in the performance of Black children, some of the deficit shown by this group may come from effects characterizing inner cities or rural areas and some from the effect of lower parental education. Thus, the observed race difference without adjustment on type of community and parental education may be spurious due to the effects of those factors.

To illustrate the problem further, let the hypothetical figures in Table 1 represent the true probability of success (i.e., probability of a student being able to pass an achievement test) in each of several subpopulations. From this table, two observations are clear: (1) Blacks and Whites from the same type of community do not differ in their achievement, and (2) children from different communities do differ in their achievement. If researchers are interested in making overall comparisons between Blacks and Whites or among types of communities, their conclusions should be consistent with the above observations.

Insert Table 1 here.

We can illustrate the difficulties which may arise in making group comparisons by considering some extreme cases. Suppose that in a survey our Black sample consisted of 100 urban children while our White sample consisted of 100 suburban children. Using the probabilities in Table 1, we would expect all the Whites to pass the test and only half the Blacks to pass it. This could represent a real race difference, but it could also be accurately described as a difference in type of community. The two factors are completely confounded. Using these data alone, it would, in fact, be impossible to say whether the difference is really a race difference or a type of community difference. If our population happened to consist of 100 Black suburban children and 100 White urban children, the performance results would be expected to turn out just the opposite; i.e., favoring Blacks. These are the two extreme cases of results one might obtain given the probabilities in Table 1. For varying population sizes, we may obtain various intermediate results; for example, given the representative sample sizes in Table 2, we have the expected number of successes in Table 3 which gives proportions of success of .7 for both Whites and Blacks. This is consistent with Table 1 to the extent that no race effect is observed.

Insert Table 2 and Table 3 here.

For further illustration, suppose that a representative sampling gives sample sizes as shown in Table 4 and the expected number of successes as shown in Table 5. The total expected number of successes for Blacks is

80 (i.e., 50 + 0 + 30), and the simple proportion of successes for Blacks regardless of the type of community is .533 (i.e., 80/150--the ratio of the total number of successes for Blacks to the total number of observations for Blacks). Likewise, the success rate for Whites is (50 + 200 + 90)/450 or .756.

Insert Table 4 and Table 5 here.

We have then an apparent difference in achievement between Blacks and Whites despite the fact that there is no such difference within any of the type of community classes in Table 1. In effect, if we adjust or control for type of community, there are no race effects. An alternative explanation for the race differences is that the two groups, Whites and Blacks, are exposed to different types of communities.

In summary, then, the use of marginal values in such tables as these will yield misleading interpretations unless the numbers of observations in the various classifications are equal or certain conditions of proportionality exist between the cell sizes. This is an unnecessary and, in some cases, an impossible restriction to meet, and there are other analytical methods of dealing with the problem of disproportionality.

B. Balancing Solution

There may be a number of ways to solve the problem of estimating and interpreting group differences in studies with disproportionate populations. One of them is called balancing, a statistical method developed in conjunction with the National Assessment of Educational Progress (NAEP). This method simultaneously "balances," for each group and category, the disproportionate representation of the other groups

that exists in the population. That is, the effects of one factor are estimated in the absence of the effects of the other factors. This method is illustrated in NAEP Report 7 (NAEP, 1971) and an invitational presentation (Ahman, et al., 1973). A brief illustration may present some flavor of the method.

Let us take the figures in Table 4 and Table 5 for illustration. To calculate the balanced or adjusted group differences, the balancing method assumes that (1) there is a common constant, denoted by $\bar{p}$; and (2) there is a unique but unknown effect attributed to each category, denoted by a_1, a_2, a_3, b_1, and b_2 for urban, suburban, rural, Black, and White, respectively. With these assumptions, each observed cell proportion can be partitioned into the common constant and unique effects. For example, $cell_{11}$ proportion can be written as $\bar{p} + a_1 + b_1$, and $cell_{12}$ proportion can be written as $\bar{p} + a_1 + b_2$. The total number of successes for the urban group can thus be written as $100(\bar{p} + a_1 + b_1) + 100(\bar{p} + a_1 + b_2) = 100$. The constant is directly obtained as the ratio of the total number of successes to the total number of observations. That is, 420/600 = .70. The other five unknowns are obtained by solving the following set of simultaneous equations. The numbers on the right side of the first five equations are simple sums across other categories. For example, 100 is the toal number of successes for urban children, which is the simple sum of 50 Blacks and 50 Whites.

$$100\ (\bar{p} + a_1 + b_1) + 100\ (\bar{p} + a_1 + b_2) = 100$$

$$0\ (\bar{p} + a_2 + b_1) + 200\ (\bar{p} + a_2 + b_2) = 200$$

$$50\ (\bar{p} + a_3 + b_1) + 150\ (\bar{p} + a_3 + b_2) = 120$$

$$100\ (\bar{p} + a_1 + b_1) + 0 + 50\ (\bar{p} + a_3 + b_1) = 80$$

$$100\ (\bar{p} + a_1 + b_2) + 200\ (\bar{p} + a_2 + b_2) + 150\ (\bar{p} + a_3 + b_2) = 340$$

$200\ (a_1 + a_2 + a_3) = 0$

$150\ b_1 + 450\ b_2 = 0$

The last two equations are the constraints. Solving these equations will yield the following results:

$a_1 = -.20 \qquad b_1 = .00$

$a_2 = +.30 \qquad b_2 = .00$

$a_3 = -.10$

These are the solutions that one would expect if the population was equally distributed or proportionately distributed across the two factors. The effects for each factor are adjusted for the effects of the other factors. It is suggested that balancing solutions be applied to marginal estimates whenever the samples or populations are disproportionately distributed across the various categories.

C. Method Equivalent to Balancing

As pointed out by Appelbaum and Cramer (1974), the balancing solution is equivalent to the least-square solution of an additive ANOVA model. If the previous data (in Tables 4 and 5) are coded in a binary fashion (i.e., 0 being failure, and 1 being success), non-orthogonal ANOVA will provide exactly the same results as the balancing method. The equivalences of balancing to ANOVA model are explicated in Appelbaum and Cramer's paper. The relation of balancing to ANOVA can be briefly described as follows:

Using the example presented in Tables 4 and 5, the additive ANOVA model for the cell mean $\bar{y}_{ij}$ can be denoted as

$$\bar{y}_{ij} = u + \alpha_i + \beta_j + \varepsilon_{ij} \ ,$$

where

u is the grand mean;

α_i are row effects corresponding to the type of community; and

β_j are column effects corresponding to race.

The constraints are

$$\sum_j n_{i.}\,\alpha_i = \sum_j n_{.j}\,\beta_j = 0$$

where

$$n_{i.} = \sum_j n_{ij} \quad \text{and} \quad n_{.j} = \sum_i n_{ij} .$$

For the total data set, the model can be partitioned into a model matrix A, a parameter vector θ^*, and an error component vector E. That is,

$$\begin{bmatrix} \bar{y}_{11} \\ \bar{y}_{12} \\ \bar{y}_{21} \\ \bar{y}_{22} \\ \bar{y}_{31} \\ \bar{y}_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} u \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{bmatrix}$$

$$\bar{y} = A\theta^* + E$$

Matrix A is a singular matrix. Thus, θ^* are not estimable. However, we can choose m linear combinations of the parameters, which can be uniquely estimated (m must be equal or less than the rank of A). The weights defining m linear combinations of parameters from the rows of a contrast matrix L. The original model can then be formulated as

$$\bar{y} = KL\theta^* + E = K\theta + E .$$

The least-squares estimate for θ is given as

$$\hat{\theta} = (K'DK)^{-1} K'D\bar{y} ,$$

where

D is the diagonal matrix of cell frequencies.

If the L matrix is defined as

$$L = \begin{bmatrix} 1 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} & 0 & 0 \\ 0 & -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & -\frac{1}{2} \end{bmatrix}$$

or in other words, if θ is defined as

$$\theta = \begin{bmatrix} u + \alpha. + \beta. \\ \alpha_1 - \alpha. \\ \alpha_2 - \alpha. \\ \beta_1 - \beta. \end{bmatrix}$$

The resulting four estimates in the order of the rows are equivalent to $\bar{p}$, a_1, a_2, and b_1, respectively. Consequently, a_3 and b_2 can be obtained. The above L matrix is a deviation contrast matrix, which can be formulated for any factorial design. In other words, balanced effects for multiple factors can be obtained with the above ANOVA estimation procedures.

What this tells us is that the balanced estimates are main effects adjusted for other main effects. Thus, balancing solutions can be obtained easily by analysis of variance using a linear model. Any

comprehensive computer program such as MANOVA (Clyde, 1969) and Multivariance (Finn, 1972) can do exactly the same kind of balancing as used by the NAEP.

One thing, however, that should be noted is that the balancing method does not incorporate interactions. The assumption of no interactions is always assumed when the balancing method is applied. If this assumption is not justifiable, balanced results may not be very meaningful because the main effects are confounded with interactions. A safeguard is to first conduct ANOVA to test for interactions. If the interactions are not statistically significant, one can then proceed to obtain the balanced solutions with an additive model. If there are interactions, other alternatives such as simple effect investigations can be adopted.

D. Application to Continuous Criterion Variables

It has been shown that, for binary data, balanced results generally are extremely useful in interpreting group differences. However, the application of balancing should not be limited to the analysis of a contingency table. The rationale of adjustment should also be applied to studies involving variables with interval properties. As mentioned previously, estimates from balancing are equivalent to those obtained from non-orthogonal ANOVA, and thus balanced estimates are estimated main effects "adjusted" for other main effects. In other words, the balanced marginal values are the so-called adjusted marginal means. Since many studies use continuous rather than binary criterion variables, the application of the balancing method to continuous variables is a straightforward generalization.

To illustrate its application to the case of a continuous dependent variable, a set of scores in the area of vocabulary from a statewide

educational assessment were analyzed with a four-factor design. The four factors are:

(1) Region - designated as A, B, and C;

(2) Type of Community - inner city, suburban, and rural;

(3) Race - Black, White, and other; and

(4) Sex.

The sample distribution is presented in Table 6. It is easily seen that the subpopulations are disproportionate (the sample is a statewide probability sample). Observed differences among regions, between males and females, or among races may be masqueraded by the effects of other factors. Thus, to obtain "better" estimates of group effects, balancing was performed.

Insert Table 6 here.

Multivariance (Finn, 1972) was used to obtain the adjusted marginal means assuming the absence of interactions. The model used was a main effects non-orthogonal ANOVA model. With this program, each set of means was balanced or adjusted by effects of the other three factors. The final results are presented in Table 7.

Insert Table 7 here.

It is seen that the adjusted estimates of group means are different from the unadjusted ones. For some groups, the difference between adjusted means becomes larger (e.g., males vs females). For other groups, the differences are narrowed after adjustment (e.g., regions). The magnitude of differences after adjustment varies from group to group. For instance, in

the absence of other effects, the balanced differences between Blacks and Whites and between Whites and others become smaller, whereas the difference between Blacks and others increases. These differences in results also show that the adjustment is not always in the direction of making group differences smaller.

E. Summary and Discussion

Because of disporportionate populations, the observed simple group effects are often masqueraded by effects of other known as well as unknown factors. This masquerading effect may make the observed scores greater or smaller than would be expected if we adjusted for disproportionality and would lead researchers to erroneous conclusions. To deal with the masquerading effect, a balancing method was proposed by NAEP to estimate proportions of students passing a certain criterion test. It has been demonstrated that this method yields estimates of group effects in the absence of effects of other known factors. It is thus suggested that balancing solutions should be applied to marginal estimates whenever the populations are disproportionate.

The balancing method is equivalent to unequal-N ANOVA of a contingency table. If data are coded in a binary fashion, ANOVA with an additive model (without interaction terms) provides exactly the same results as the balancing method. This indicates that ANOVA can be used to perform balancing.

While the original balancing method is geared to binary data, the idea of adjustment can be generalized to data of interval variables. Since balancing is equivalent to non-orthogonal ANOVA, the balanced group effects are the so-called adjusted main effects. In studies in which interval

variables are used, the adjusted marginal means or adjusted main effects should be used to describe group status and group differences.

The state assessment example in this paper further accentuates the likelihood of coming to erroneous conclusions based upon observed (unadjusted) scores. For instance, the difference between Regions A and C is reduced from the observed 6.77 to the balanced 2.25 on vocabulary achievement. If the effects are not adjusted, the results are misleading and could lead to erroneous decisions.

While balancing, as proposed by NAEP, appears to be useful in most situations, its assumption of no interaction may be too restrictive in some cases. Since non-orthogonal ANOVA is an equivalent method, and is capable of testing for the existence of interaction, it should be applied to examine the possible interaction first before performing the balancing.

It should be noted that balanced results cannot be used as a basis for drawing causal inferences unless the study is an experimental one in which subjects were randomly assigned to the "treatment" groups. There are many problems in interpreting balanced results. For example, some independent variables that should be used for balancing may be unavailable or even unknown while some balancing variables may be poorly measured.

Consequently, the selection and measurement of balancing variables is critical for the proper estimation of balanced effects. It is obvious that the independent (classification) variables must be related to outcome measures. Literature research, field test, and relevant theories may help to determine the possible relationship between the independent variables and outcome measures. The independent variables should also be well defined and measured with valid and reliable instruments. For example,

SES and aptitude are not as well defined as sex and race. Before using these variables, considerations must be given to proper measurement procedures. Failure to use properly measured independent variables may further complicate the balanced results and, in fact, may lead the investigator to erroneous conclusions.

It should also be noted that balancing solutions do not always reduce or increase group differences. It is not intended to make group differences look better or worse. It simply attempts to obtain "better" estimates of group differences.

In summary, when the variables used for balancing are properly selected and measured, the balanced results probably come closer to the true situation in contrast to unbalanced ones.

References

Ahman, J. S., et al. A look at the analysis of national assessment data. In W. E. Coffman, Ed., Frontiers of educational assessment and information systems--1973. Boston: Houghton Mifflin Co., 1973, 89-111.

Appelbaum, M. I., & Cramer, E. M. Balancing--the equivalence of methods. L. L. Thurstone Psychometric Laboratory Report Series, No. 134, 1974.

Clyde, D. J. Multivariate analysis of variance on large computers. Miami: Clyde Computing Service, 1969.

Fir , J. D. Multivariance: univariate and multivariate analysis of variance, covariance, and regression. Ann Arbor: National Education Resources, Inc., 1972.

National Assessment of Educational Progress, Report 7. Group and balanced group results for color, parental education, size and type of community and balanced group results for region of the country, sex. Science, December, 1971.

1/ This work was supported in part by Grant NE-G-00-3-0111 from the National Institute of Education, U. S. Department of Health, Education, and Welfare to the Psychometric Laboratory, University of North Carolina at Chapel Hill. The opinions expressed do not necessarily reflect the position or policy of the National Institute of Education and no official endorsement by the National Institute of Education should be inferred.

Table 1

True Probability of Success

Type of Community	Black	White
Urban	.50	.50
Suburban	1.00	1.00
Rural	.60	.60

Table 2

Representative Sample Sizes

	Black	White	Total
Urban	100	100	200
Suburban	100	100	200
Rural	100	100	200
Total	300	300	600

Table 3

Expected Number of Successes

	Black	White	Total
Urban	50	50	100
Suburban	100	100	200
Rural	60	60	120
Total	210	210	420
Overall P	.70	.70	.70

Table 4

Representative Sample Size

	Black	White	Total
Urban	100	100	200
Suburban	0	200	200
Rural	50	150	200
Total	150	450	600

Table 5

Expected Number of Successes

	Black	White
Urban	50	50
Suburban	0	200
Rural	30	90

Table 6

Sizes of Samples

Region	Sex	Urban			Suburban			Rural		
		Black	White	Other	Black	White	Other	Black	White	Other
A	M	15	23	0	35	112	1	156	136	18
	F	13	24	0	36	78	1	144	108	13
B	M	64	164	2	57	179	0	65	203	1
	F	100	126	0	75	174	0	60	187	0
C	M	2	6	0	4	50	0	5	130	0
	F	5	7	0	6	55	0	8	122	1

Table 7

Observed (and Adjusted) Means for Each Subgroup

	N	Mean		N	Mean
Males	1428	80.08 (77.68)	Urban	551	84.22 (81.83)
Females	1343	82.27 (80.29)	Suburban	863	82.39 (78.65)
			Rural	1357	79.10 (76.49)
Region A	913	77.94 (77.99)	Black	850	73.04 (74.23)
Region B	1457	82.16 (78.73)	White	1884	84.94 (85.94)
Region C	401	84.71 (80.24)	Other	37	73.57 (76.79)